Finding SNPs in a Sequence Using BLAST

Exploring the Sequence Viewer display of BLAST results for SNP discovery

https://blast.ncbi.nlm.nih.gov & https://www.ncbi.nlm.nih.gov/snp/

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Background

NCBI maintains a large public repository for human Short Nucleotide Variations database, also known as dbSNP. It can be accessed through the web from its homepage at www.ncbi.nlm.nih.gov/snp. These variation data are fully integrated with other NCBI sequence datasets, as well as its flagship sequence alignment tool - BLAST sequence alignment service at blast.ncbi.nlm.nih.gov.

NCBI used to provide a BLAST service to search against the flanking sequences of existing SNVs. Using this service in variation identification was cumbersome for several reasons:

- Variations identified using new technologies do not provide their true flanking sequences
- Variations identified from array experiments have very short flanking sequences, some of which could map to multiple genomic locations
- Long flanking sequences for some variations often have ambiguous base calls making variant calls and their mapping difficult
- Insertion and deletion alleles are generally represented by an N at the FASTA sequence level, which makes BLAST searches and their interpretation difficult without referencing the allele in the FASTA definition line
- Lastly, this approach did not provide the important genomic context and coordinates of identified variations

However, many reported human nucleotide variations in the biomedical literature often uses allele plus some flanking sequences to represent the variations, this makes correlating the all important phenotype and disease impact information to existing rsIDs in dbSNP as well as examine them in the proper genomic context difficult to achieve.

In this tutorial, we will demonstrate a way to map variants, contained within example sequences, using NCBI BLAST tool, through examining the genomic or mRNA alignment through the Graphical Sequence Viewer (SV) [1] linked embedded in the BLAST report [2], as well as interactive manipulating the display to better present the results in the context of gene features and SNV annotation.

Practical applications

Researchers often attempt to identify variations in specific genes or regions of the genome. *In silico* analysis of specific sequences by alignment to those in public collections can identify patterns of mismatches, which could represent a potential variation. Mapping variations to existing records in dbSNP can help connect them to functional analyses reported in literature and facilitate decision-making processes to enhance the speed and productivity of research. For example, two query sequences, a genomic sequence fragment GQ892012.1 and a cDNA clone CK130262.1, are related to the human IL4 gene. According to published articles (go.usa.gov/xGXRK), they may contain variations that have been shown to affect the disease prognosis. In the following sections of this handout, we will demonstrate how to identify sequence variations contained in these two sequence records through BLAST alignment and match them to highly informative records in dbSNP.

Setting up the BLAST searches

Selection of BLAST algorithm

To identify dbSNP records that map to the query sequence, select “nucleotide blast” from the “Web BLAST” section on the BLAST homepage (blast.ncbi.nlm.nih.gov).

Query input

The BLAST search pages support queries in raw sequence, sequences in FASTA format, or identifiers of sequence (GIs or accessions). Accession numbers GQ892012.1 or CK130262.1 will be used in the Query box in this set of exercises.

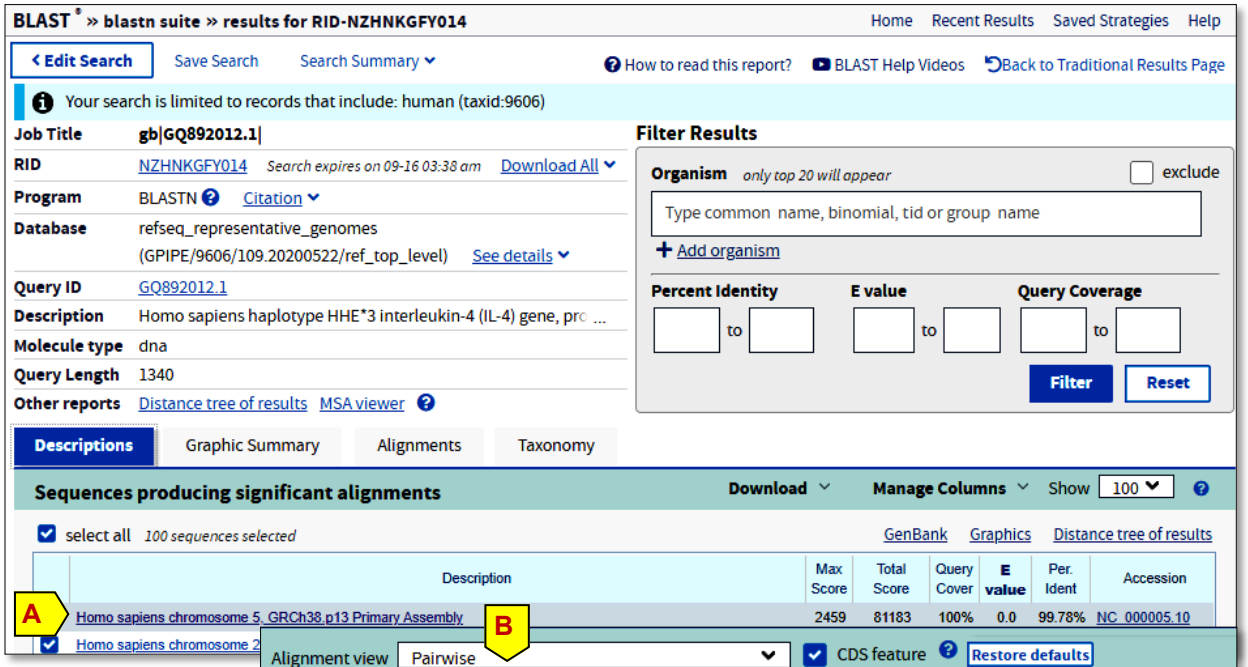
Database setup

Different types of input sequences require different databases. With genomic DNA sequences as the input query, select “RefSeq Representative genomes” as the target database. To search with expressed sequences, select “Reference RNA sequences (refseq_rna)” instead. You can further restrict database lookup to a particular organism using the Organism input box. This limit speeds up the search, makes the results returned more focused, and greatly simplifies the subsequent analyses. Submit the search by clicking the “BLAST” button. The web page automatically checks for results during the search, and renders the results in the browser window when the search is completed.

For your convenience, pre-configured BLAST search pages for a human cDNA and genomic query are listed below:
go.usa.gov/xGX9Q for the genomic query (GQ892012.1), and
go.usa.gov/xGXMn for the cDNA query (CK130262.1)

Examination of the BLAST result

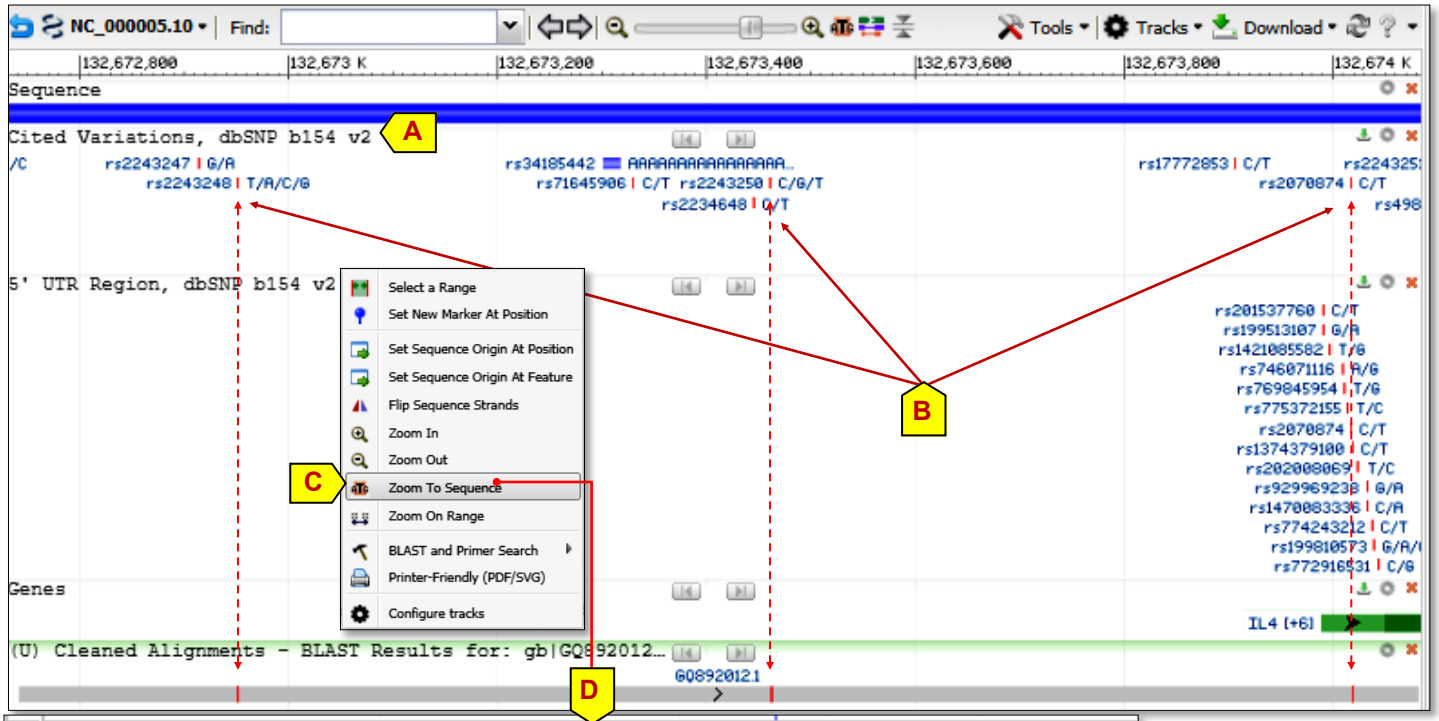
The genomic BLAST search result for accession GQ892012.1, preserved under RID NZHNKGFY014, is shown to the right. The top hit in the Descriptions table is to a region on Chromosome 5 from the GRCh38.p13 Primary Assembly (**A**). Clicking the title of that entry changes the display to the corresponding alignment (**B**) for more details. A mismatch in the alignment indicates the presence of a potential variation (circled), about 1.1 kb upstream the translation start site (**C**).

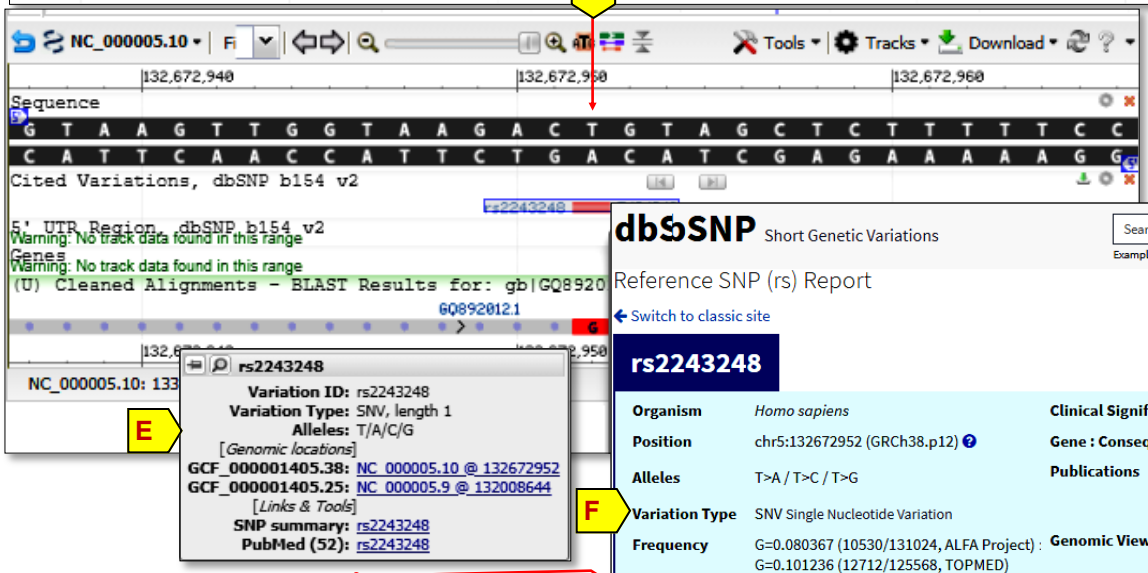


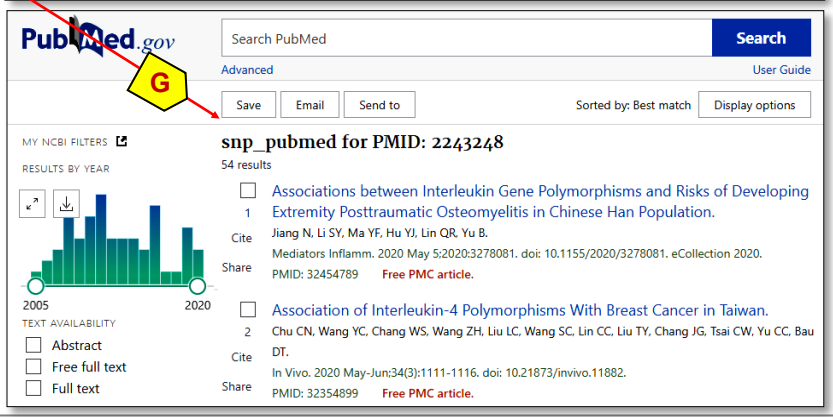
The Graphics link (**D**) displays the alignment in SV and allows the visualization of the mismatch under the context of existing annotation. This SV display showing the alignment between the query GQ892012.1 and a region from the GRCh38.p13 Primary Assembly, where mismatches are represented by red tick marks (**E**). The display is customized with annotated variation tracks (**F**), added using the **Tracks** dialog box (**G**).

Visualizing SNP-related tracks for a genome sequence

The addition Cited Variations tracks to the graphical display (below) shows all mapped variations with association publications (**A**). All three mismatches in this alignment appear to correlate with annotated variations that are referenced in publications (**B**). Right-click a region around a desired variant to see the context menu, select the "Zoom To Sequences" option (**C**) to see more details at the sequence level.





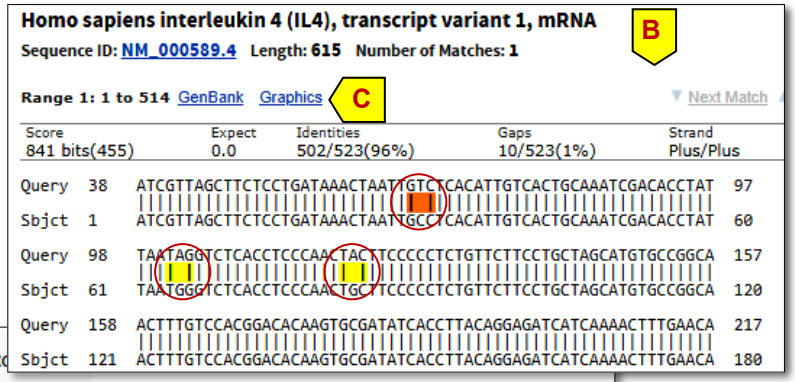


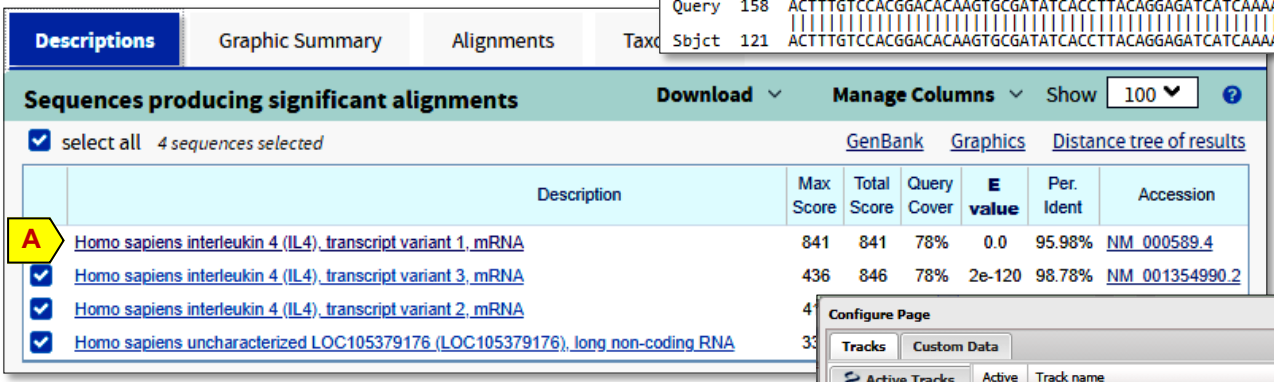
In the sequence-level view (**D**), we can see that the specific change the input query has at this mismatched position is G, vs the T in the reference genomic sequence.

Hovering over a variation activates a popup box with additional details about the variation (**E**). Clicking the rsID opens the SNP records in a new window (**F**). The PubMed link (**G**) points to the set of publications with research findings on this specific SNP record.

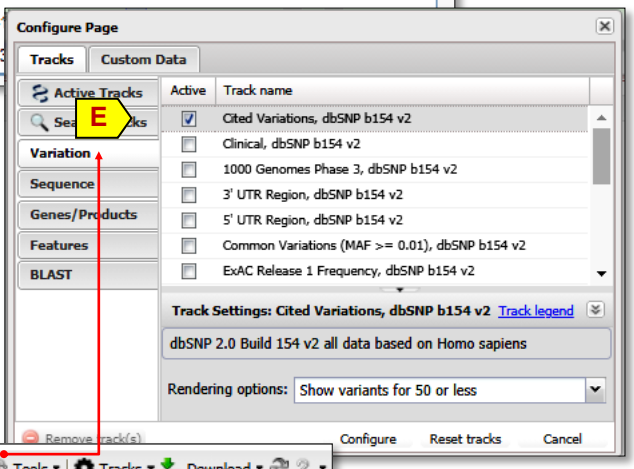
SNP mapping to RefSeq mRNAs

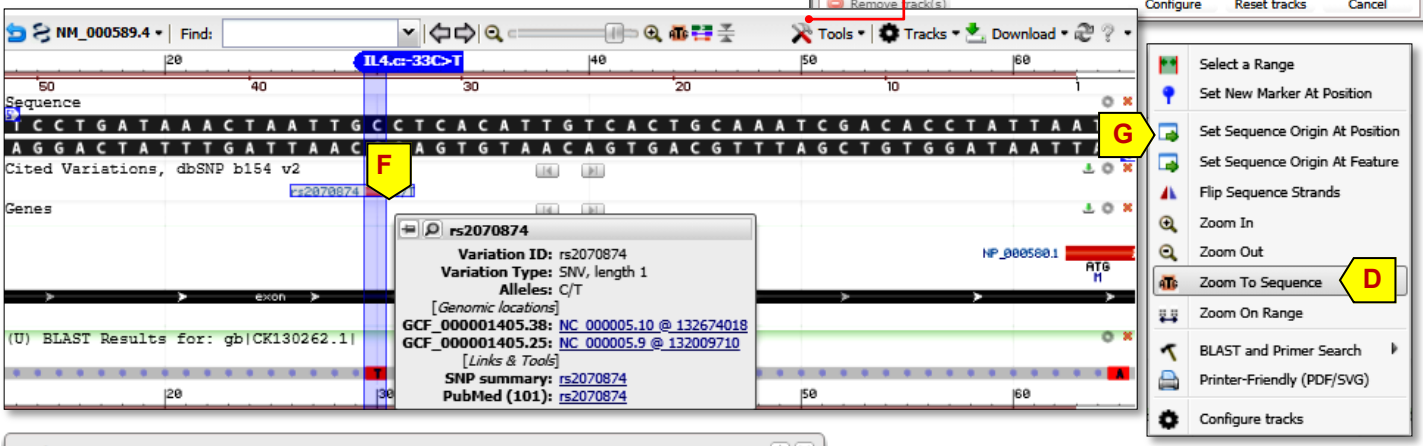
The RefSeq RNA BLAST result for CK130262.1, preserved under RID P09WMZYP01R, is shown below and to the right. The top hit in the Descriptions table is the IL4 transcript variant NM_000589.4 (**A**). Clicking the record title change the page to display the alignment (**B**) where three mismatches (circled) are shown. Clicking the Graphics link (**C**) displays the alignment in SV to allow interactive examination of variations in the query under the context of NCBI annotations.

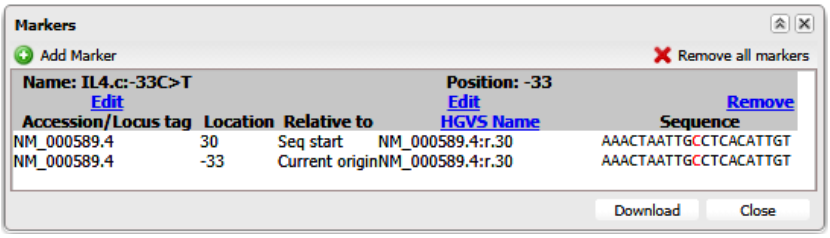




Customizing the SV display around the first mismatch by zooming to the sequence level using context menu option (**D**) and adding the Cited Variation track under Variation (**E**), it is clear that the first mismatch corresponds with the existing variation record rs2070874 (**F**). To better identify the variations using the naming convention from the Human Genome Variation Society (HGVS), you can locate the position of A for the start codon, right-click and select "Set Sequence Origin At Position" (**G**) to reset the numbering of this mRNA to begin at the start of the open reading frame. The adjusted position of the variant is 33 bases upstream from the start of the coding start.







The "Marker Details" option in the marker context menu (not shown) displays a table (**H**) to provide the location relative to the newly set sequence origin. This takes manual counting out of the process, and the information put the HGVS annotation of this variant as NM_000589.4:c.-33C>T.

References

1. The Graphical Sequence Viewer: ftp.ncbi.nih.gov/pub/factsheets/Factsheet_Graphical_SV.pdf
2. The New BLAST Results Page: ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewResultPage.pdf
3. SNP: Database for Short Genetic Variations: ftp.ncbi.nlm.nih.gov/pub/factsheets/Factsheet_SNP.pdf